

# StackIQ Enterprise Data Reference Architecture



StackIQ and Hortonworks worked together to Bring You World-class Reference Configurations for Apache Hadoop Clusters.

## Contents

The Need for Efficiency and Automation in Hadoop Deployments 2

StackIQ Enterprise Data 3

Key Benefits of StackIQ Enterprise Data 6

Enterprise Hadoop Use Cases 6

StackIQ Enterprise Data Reference Architecture 8

Summary 12

For More Information 12

About StackIQ 13

## Abstract

*As Big Data applications and the Big Infrastructure to support them have grown in popularity, the complexity of managing these solutions has become more apparent. Nowhere is this more apparent than with Apache Hadoop, the leading software framework for Big Data applications. Until now, one team of administrators has been responsible for installing and configuring cluster hardware, networking components, software, and middleware in the foundation of a Hadoop cluster. Another team has been responsible for deploying and managing Hadoop software on the cluster infrastructure. These tasks have relied on a variety of tools – both new and legacy – to handle deployment, scalability, and management. Most of the tools require management of changes to configurations and other fixes through the writing of software scripts. Making changes to one appliance or several appliances entails a manual, time-intensive process that leaves the administrator uncertain as to whether the changes have been implemented throughout the application cluster. Using homogenous, robust hardware/software solutions for Big Data applications from EMC, Teradata, Oracle, and others is another, very expensive, and more limited, alternative.*

*Now, however, a paradigm shift in the design, deployment, and management of Big Data applications is underway. For the first time in the industry, a best-of-breed Hadoop distribution has been combined with best-of-breed Hadoop and cluster management software. The result is StackIQ Enterprise Data – a complete enterprise solution. Cost, efficiency, agility, and reliability make StackIQ Enterprise Data unique. This whitepaper describes our reference architecture for Hadoop deployments.*

## The Need for Efficiency and Automation in Hadoop Deployments


The Apache Hadoop software framework has become the leading solution for massive, data-intensive, distributed applications. More mature than other solutions, it has also proven to be better at scaling; more useful, flexible, and affordable as a generic rather than proprietary data platform; excellent at handling structured and unstructured data; and its many connector products have broadened its use beyond other software frameworks used to handle Big Data applications.

The growing popularity of Hadoop has also put a spotlight on its shortcomings – specifically, the complexity of deploying and managing Hadoop infrastructure. Early adopters of Hadoop have found that they lack the tools, processes, and procedures to deploy and manage it efficiently. IT organizations are facing challenges in coordinating the rich clustered infrastructure necessary for enterprise-grade Hadoop deployments.

The current generation of Hadoop products was designed for IT environments in which different groups of skilled personnel are required to deploy them. One group of IT professionals installs and configures the cluster hardware, networking components, software, and middleware that form the foundation of a Hadoop cluster. Another group is responsible for deploying the Hadoop software as part of the cluster infrastructure. Most cluster management products focus on the upper layers of the cluster (i.e. Hadoop products, including the Hadoop Distributed File System, MapReduce, Pig, Hive, HBase, and Zookeeper), leaving the installation and maintenance of the underlying server cluster to other solutions. Thus the overall Hadoop infrastructure is deployed and managed by a collection of disparate products, policies, and procedures, which can lead to unpredictable and unreliable clusters.

Combining the leading Apache Hadoop software stack with the leading cluster management solution, StackIQ has engineered a revolutionary new solution that

“StackIQ has engineered a revolutionary new solution.”



makes Hadoop deployments of all sizes much faster, less costly, more reliable, and more flexible. StackIQ Enterprise Data optimizes and automates the deployment and management of underlying cluster infrastructures of any size while also providing a massively scalable, open source Hadoop platform for storing, processing, and analyzing large data volumes.

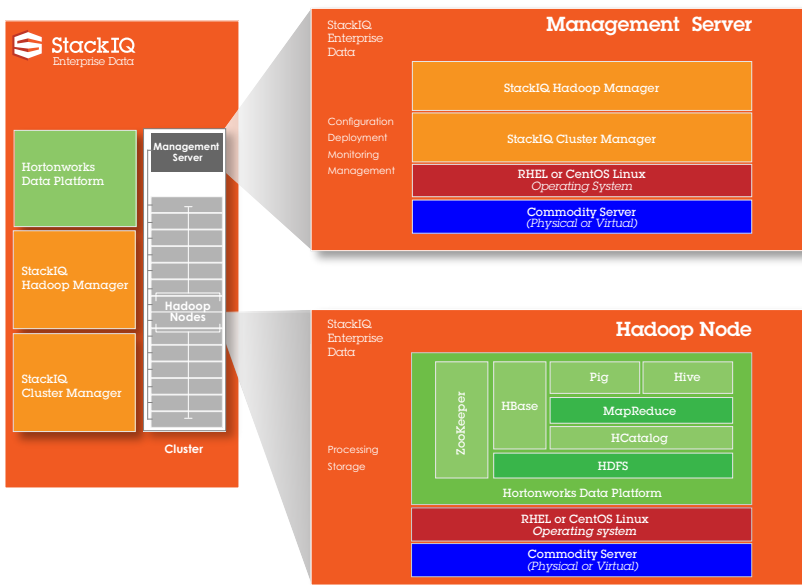
With StackIQ Enterprise Data, physical or virtual Hadoop clusters can be quickly provisioned, deployed, monitored, and managed. System administrators can manage the entire system using a single pane of glass. New nodes are also configured automatically from bare metal—with a single command—without the need for complex administrator assistance. If a node needs an update, it will be completely re-provisioned by the system to ensure it boots into a known good state. Since StackIQ Enterprise Data places every bit on every node, administrators have complete control and consistency across the entire infrastructure. Now administrators have the integrated, holistic Hadoop tools and control they need to more easily and swiftly meet their enterprise Big Data application requirements.

### **StackIQ Enterprise Data**

StackIQ Enterprise Data is a complete, integrated Hadoop solution for enterprise customers. For the first time, enterprises get everything they need to deploy and manage Hadoop clusters throughout the entire operational lifecycle in one product (Figure 1). StackIQ Enterprise Data includes:

**Hortonworks Data Platform** powered by Apache Hadoop is an open-source, massively scalable, highly stable and extensible platform based on the most popular and essential Hadoop projects for storing, processing, and analyzing large volumes of structured and unstructured data. Hortonworks Data Platform platform makes it easier than ever to integrate Apache Hadoop into existing data architectures. Highly recommended for anyone who has encountered difficulties installing and integrating Hadoop projects downloaded directly from Apache, Hortonworks Data Platform is also ideal for solution providers wanting to integrate or extend their solutions for Apache Hadoop.

The platform includes HDFS, MapReduce, Pig, Hive, HBase, and Zookeeper, along with open source technologies that make the Hadoop platform more manageable, open, and extensible. These include HCatalog, a metadata management service for simplifying data sharing between Hadoop and other enterprise information systems, and a complete set of open APIs such as WebHDFS to make it easier for ISVs to integrate and extend Hadoop.



Hortonworks has contributed more than 80% of the code in Apache Hadoop to date and is the main driving force behind the next generation of the software. The team has supported the world’s largest Hadoop deployment, featuring more than 42,000 servers. Competitive products offer altered, non-standard versions of Hadoop, often complicating integration with other systems and data sources. Hortonworks is the only platform that is completely consistent with the open source version.

Figure 1. StackIQ Enterprise Data Components

**StackIQ Hadoop Manager** manages the day-to-day operation of the Hadoop software running in the clusters, including configuring, launching, and monitoring HDFS, MapReduce, ZooKeeper, Hbase and Hive. A unified single pane of glass—with a command line interface (CLI) or graphical user interface (GUI)—is used to control and monitor all of these, as well as manage the infrastructure components in the cluster.

Easy to use, the StackIQ Hadoop Manager allows for the deployment of Hadoop clusters of all shapes and sizes (including heterogeneous hardware support, parallel disk formatting, and multi-distribution support). Typically, the installation and management of a Hadoop cluster has required a long, manual process. The end user or deployment team has had to install and configure each component of the software stack by hand, causing the setup time for such systems and the ongoing

management to be problematic and time-intensive with security and reliability implications. StackIQ Enterprise Data completely automates the process.

**StackIQ Cluster Manager** manages all of the software that sits between bare metal and a cluster application, such as Hadoop. A dynamic database contains all of the configuration parameters for an entire cluster. This database is used to drive machine configuration, software deployment (using a unique Avalanche peer-to-peer installer), management, and monitoring. Regarding specific features, the Cluster Manager:

- Provisions and manages the operating system from bare metal, capturing networking information (such as MAC addresses)
- Configures host-based network settings throughout the cluster
- Captures hardware resource information (such as CPU and memory information) and uses this information to set cluster application parameters
- Captures disk information and using this information to programmatically partition disks across the cluster
- Installs and configuring a cluster monitoring system
- Provides a unified interface (CLI and GUI) to control and monitor all of this.

The StackIQ Cluster Manager for Hadoop is based on StackIQ's open source Linux cluster provisioning and management solution, Rocks, originally developed in 2000 by researchers at the San Diego Supercomputer Center at the University of California, San Diego. Rocks was initially designed to enable end users to easily, quickly, and cost-effectively build, manage, and scale application clusters for High Performance Computing (HPC). Thousands of environments around the world now use Rocks.

In StackIQ Enterprise Data, the Cluster Manager's capabilities have been expanded to not only handle the underlying infrastructure but to also handle the day-to-day operation of the Hadoop software running in the cluster. Other competing products fail to integrate the management of the hardware cluster with the Hadoop software stack. By contrast, StackIQ Enterprise Data operates from a continually updated, dynamic database populated with site-specific information on both the underlying cluster infrastructure and running Hadoop services. The product includes everything from the operating system on up and packages CentOS Linux or Red Hat Enterprise

## Key Benefits of StackIQ Enterprise Data

- The first complete, integrated, Hadoop solution for the enterprise
- Faster time to deployment
- Automated, consistent, dependable deployment and management
- Simplified operation that can be quickly learned without systems administration experience
- Reduced downtime due to configuration errors
- Reduced total cost of ownership for Hadoop clusters

Linux, cluster management middleware, libraries, compilers, and monitoring tools.

## Enterprise Hadoop Use Cases

Hadoop enables organizations to move large volumes of complex and relational data into a single repository where raw data is always available. With its low-cost, commodity servers and storage repositories, Hadoop enables this data to be affordably stored and retrieved for a wide variety of analytic applications that can help organizations increase revenues by extracting value such as strategic insights, solutions to challenges, and ideas for new products and services. By breaking up Big Data into multiple parts, Hadoop allows for the processing and analysis of each part simultaneously on server clusters, greatly increasing the efficiency and speed of queries.

The use cases for Hadoop are many and varied, impacting disciplines as varied as public health, stock and commodities trading, sales and marketing, product development, and scientific research. For the business enterprise, Hadoop use cases include:

- **Data Processing:** Hadoop allows IT departments to extract, transform, and load (ETL) data from source systems and to transfer data stored in Hadoop to and from a database management system for the performance of advanced analytics; it is also used for the batch processing of large quantities of unstructured and semi-structured data.

- Network Management: Hadoop can be used to capture, analyze, and display data collected from servers, storage devices, and other IT hardware to allow administrators to monitor network activity and diagnose bottlenecks and other issues.
- Retail Fraud: Through monitoring, modeling, and analyzing high volumes of data from transactions and extracting features and patterns, retailers can prevent credit card account fraud.
- Recommendation Engine: Web 2.0 companies can use Hadoop to match and recommend users to one another or to products and services based on analysis of user profile and behavioral data.
- Opinion Mining: Used in conjunction with Hadoop, advanced text analytics tools analyze the unstructured text of social media and social networking posts, including Tweets and Facebook posts, to determine the user sentiment related to particular companies, brands or products; the focus of this analysis can range from the macro-level down to the individual user.
- Financial Risk Modeling: Financial firms, banks, and others use Hadoop and data warehouses for the analysis of large volumes of transactional data in order to determine risk and exposure of financial assets, prepare for potential “what-if” scenarios based on simulated market behavior, and score potential clients for risk.
- Marketing Campaign Analysis: Marketing departments across industries have long used technology to monitor and determine the effectiveness of marketing campaigns; Big Data allows marketing teams to incorporate higher volumes of increasingly granular data, like click-stream data and call detail records, to increase the accuracy of analysis.
- Customer Influencer Analysis: Social networking data can be mined to determine which customers have the most influence over others within social networks; this helps enterprises determine which are their most important and influential customers.
- Analyzing Customer Experience: Hadoop can be used to integrate data from previously siloed customer interaction channels (e.g., online chat, blogs, call centers) to gain a complete view of the customer experience; this enables enterprises to understand the impact of one customer interaction channel on another in order to optimize the entire customer lifecycle experience.
- Research and Development: Enterprises like pharmaceutical manufacturers use Hadoop to comb through enormous volumes of text-based research and other historical data to assist in the development of new products.



## Reference Architecture

Table 1 shows the StackIQ Enterprise Data reference architecture hardware using Dell PowerEdge servers.

Using 3 TB drives in 18 data nodes in a single rack, this configuration represents 648 TB of raw storage. Using HDFS's standard replication factor of 3, yields 216 TB of usable storage.

Table 1. StackIQ Enterprise Data Reference Architecture (Hardware)

Reference Hardware Configuration on Dell™ PowerEdge Servers				
Machine Function	Management Node	Name Node	Secondary Name Node	Data Node
Platform	PowerEdge R410	PowerEdge R720xd		
CPU	2 x E5620 (4 core)	2 x E5-2640 (6-core)		
RAM	16 GB	96 GB		
Network	1 x Dell PowerConnect 5524 Switch, 24-ports 1 Gb Ethernet (per rack)			
	1 x Dell PowerConnect 8024F 10Gb Ethernet switch (For rack interconnection in multi-rack configurations)			
Disk	2 x 1 TB SATA 3.5"	12 x 3TB SATA 3.5"		
Storage Controller	PERC H710			
RAID	RAID 1	NONE		
Minimum per Pod	1	1	1	3*

\* Based on HDFS's standard replication factor of 3





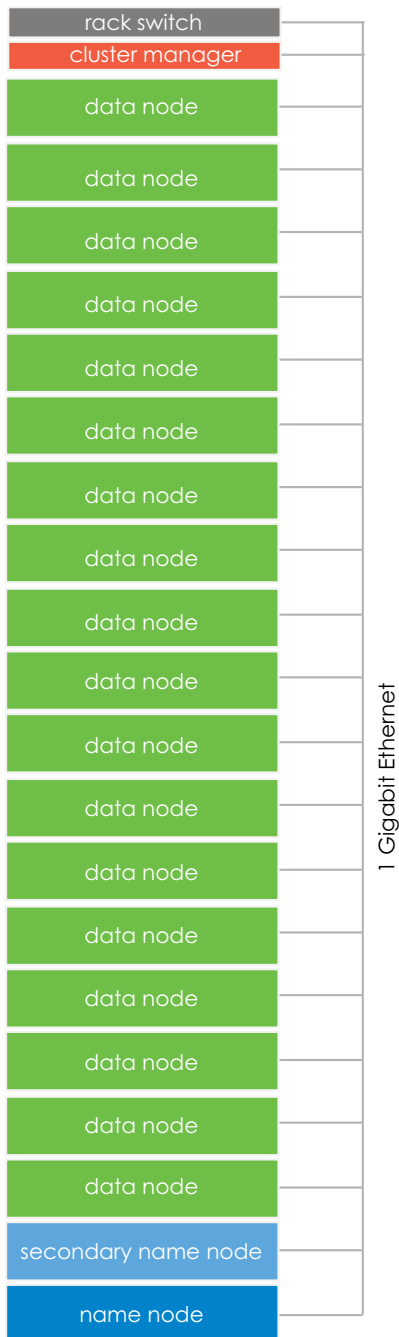
Table 2 shows the software components of the StackIQ Enterprise Data reference architecture.

The management node is installed with StackIQ Enterprise Data management software, which automatically installs and configures Hortonworks Data Platform software on the Name Node, Secondary Name Node, and all Data Nodes.

Rolls are pre-packaged software modules that integrate software components for site-specific requirements. They may be selected and automatically configured in StackIQ Enterprise Data and are available from StackIQ at <http://www.stackiq.com/download/>.

Table 2. StackIQ Enterprise Data Reference Architecture (Software)

Reference Architecture (Software)	
StackIQ Enterprise Data 1.0 ISO Contents	
Hadoop Roll	Hortonworks Data Platform 1.0
Base Roll	Rocks+ 6.0.2 Base Command Line Interface (CLI)
Kernel Roll	Installation Support for Latest x86 chipsets
Core Roll	Rocks+ 6.0.2 Core, GUI
OS Roll	CentOS 6.2
Ganglia Roll	Cluster Monitoring
Web Server Roll	Apache Web Server and WordPress



## Single Rack Configuration

In the single rack configuration, there is one Cluster Manager Node, one Name Node, and one Secondary Name Node. This configuration may include between one and 18 Data Nodes, depending upon how much storage is needed for the cluster. The top-of-rack switch connects all of the nodes using Gigabit Ethernet. A sample single-rack configuration of StackIQ Enterprise Data is shown in Figure 2.

Figure 2. Single Rack Configuration

## Multi-Rack Configuration

More racks may be added to build a multi-rack configuration. Each rack may contain between one and 20 Data Nodes, depending upon how much storage is needed for the cluster. A multiport 10 GE switch should be added to the second rack, with all of the top-of-rack switches connected to it via one of their 10 GE ports. For simplicity, a step and repeat layout is shown in the multi-rack sample configuration in Figure 3.

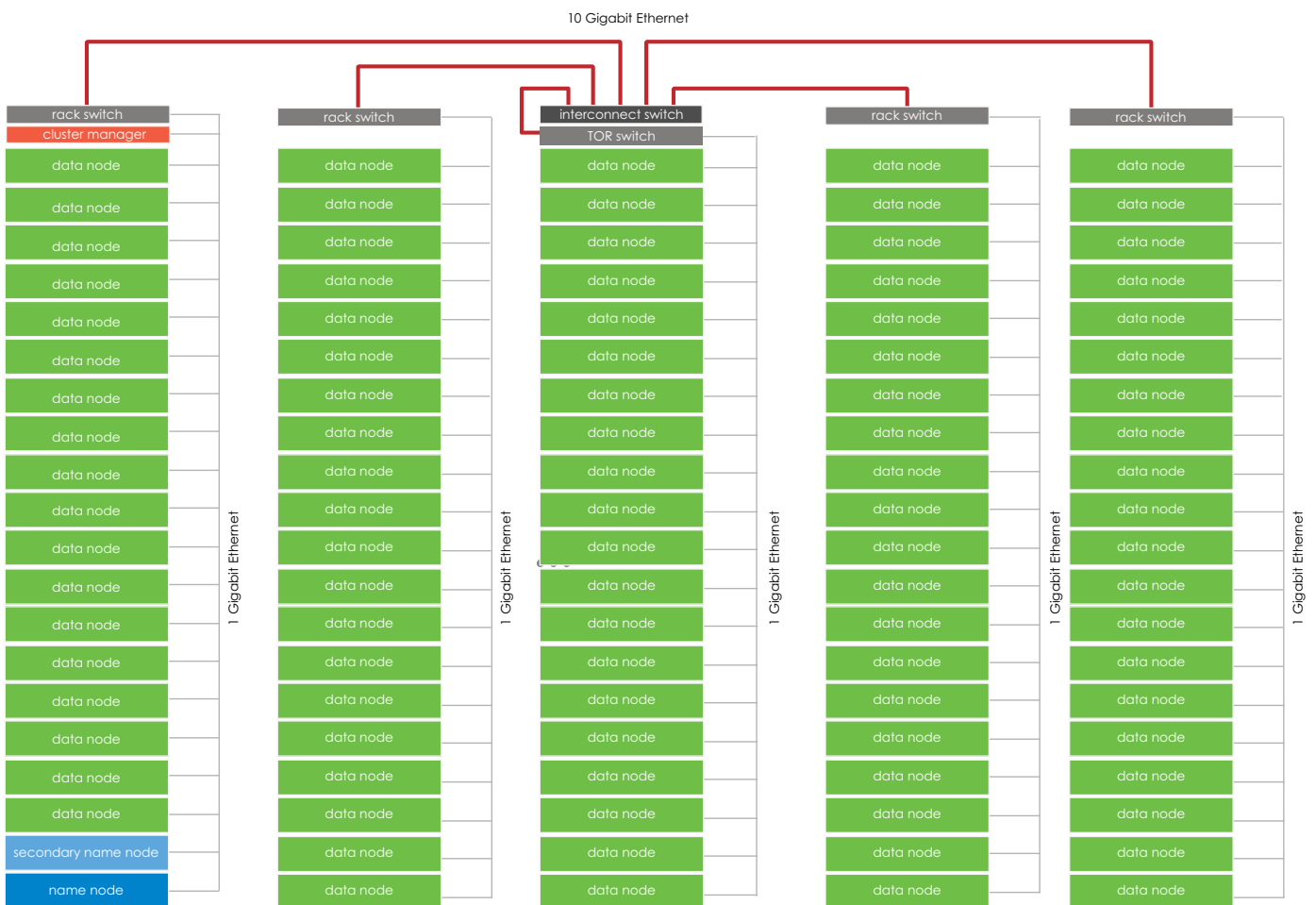


Figure 3. Multi-Rack Configuration

## Summary

As the leading software framework for massive, data-intensive, distributed applications, Apache Hadoop has gained tremendous popularity, but the complexity of deploying and managing Hadoop server clusters has become apparent. Early adopters of Hadoop moving from proofs-of-concept in labs to full-scale deployment are finding that they lack the tools, processes, and procedures to deploy and manage these systems efficiently. For reliable, predictable, simplified, automated Hadoop enterprise deployments, StackIQ has created StackIQ Enterprise Data. This powerful, holistic, simplified tool for Hadoop deployment and management combines the leading Apache Hadoop software stack with the leading cluster management solution. StackIQ Enterprise Data makes it easy to deploy and manage consistent Hadoop installations of all sizes and its automation, powerful features, and ease of use lower the total cost of ownership of Big Data systems.

## For More Information

StackIQ White Paper on “Optimizing Data Centers for Big Infrastructure Applications”  
[bit.ly/N4haal](http://bit.ly/N4haal)

Intel® Cloud Buyers Guide to Cloud Design and Deployment on Intel® Platforms  
[bit.ly/L3xXWI](http://bit.ly/L3xXWI)

Hadoop Training and Certification Programs  
[hortonworks.com/hadoop-training/](http://hortonworks.com/hadoop-training/)

Why Apache Hadoop?  
[hortonworks.com/why-hadoop/](http://hortonworks.com/why-hadoop/)

## About StackIQ

StackIQ is a leading provider of Big Infrastructure management software for clusters and clouds. Based on open-source Rocks cluster software, StackIQ's Rocks+ product simplifies the deployment and management of highly scalable systems. StackIQ is based in La Jolla, California, adjacent to the University of California, San Diego, where the open-source Rocks Group was founded. Rocks+ includes software developed by the Rocks Cluster Group at the San Diego Supercomputer Center at the University of California, San Diego, and its contributors. Rocks® is a registered trademark of the Regents of the University of California.



4225 Executive Square  
Suite 1000  
La Jolla, CA 92037  
858.380.2020  
info@stackiq.com

StackIQ and the StackIQ Logo are trademarks of StackIQ, Inc. in the U.S. and other countries. Third-party trademarks mentioned are the property of their respective owners. The use of the word partner does not imply a partnership relationship between StackIQ and any other company.